

# Canadian Bioinformatics Workshops

[www.bioinformatics.ca](http://www.bioinformatics.ca)

[bioinformaticsdotca.github.io](https://bioinformaticsdotca.github.io)

Supported by



Creative Commons

This page is available in the following languages:

Afrikaans Български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto  
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)  
Euskara Suomi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macdonian Melayu  
Nederlands Norsk Sesotho sa Leboa potski Português română slovenski jezik српски српски (latinica) Sotho svenska  
中文 華語 (台灣) isiZulu



## Attribution-Share Alike 2.5 Canada

### You are free:



to Share — to copy, distribute and transmit the work



to Remix — to adapt the work



### Under the following conditions:



**Attribution.** You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).



**Share Alike.** If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

[Disclaimer](#)

Your fair dealing and other rights are in no way affected by the above.  
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:  
[English](#) [French](#)

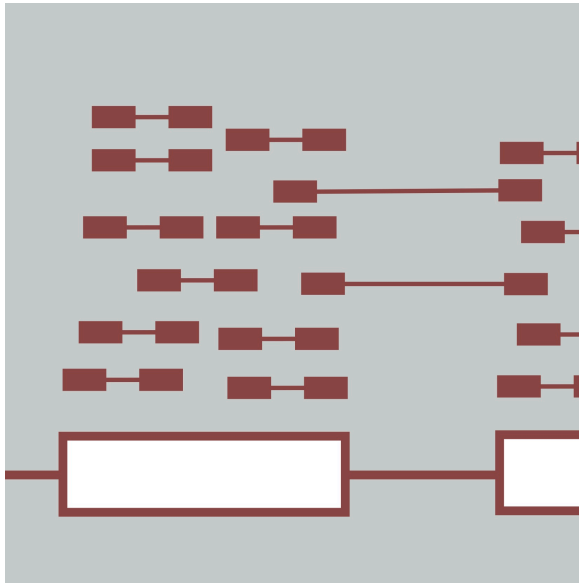
[Learn how to distribute your work using this licence](#)

# Introduction to Cloud Computing

Kelsy Cotto, Malachi Griffith, Obi Griffith, Megan Richters

Informatics for RNA-seq Analysis

June 11-13, 2019



# Learning objectives of the course

- **Module 1: Introduction to Cloud Computing**
- Module 2: Introduction to RNA Sequencing
- Module 3: Alignment and Visualization
- Module 4: Expression and Differential Expression
- Module 5: Alignment Free Expression Estimation
- Module 6: Isoform Discovery and Alternative Expression
- Module 7: Genome Free De Novo Transcript Assembly
- Module 8: Functional Annotation and Analysis of Transcripts
  
- Tutorials
  - Provide a working example of an RNA-seq analysis pipeline
  - Run in a ‘reasonable’ amount of time with modest computer resources
  - Self contained, self explanatory, portable

# Learning Objectives

- Introduction to cloud computing concepts
- Introduction to cloud computing providers
- Use the Amazon EC2 console to create an instance for each student
  - Will be used for many hands-on tutorials throughout the course
- How to log into your cloud instance

# Disk Capacity vs Sequencing Capacity, 1990-2012

Disk Storage  
(Mbytes/\$)

Stein *Genome Biology* 2010, 11:207  
<http://genomebiology.com/2010/11/5/207>

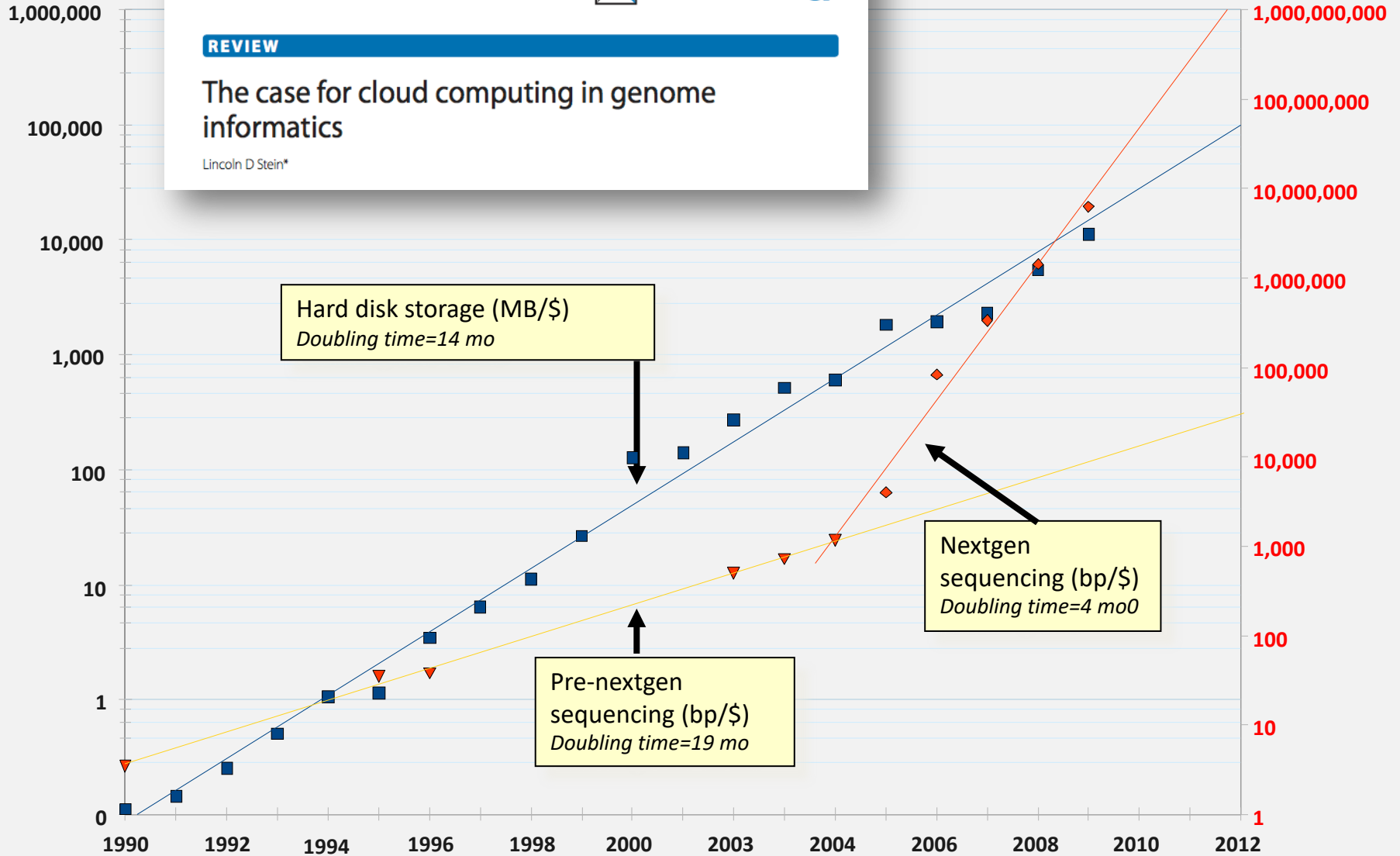


REVIEW

The case for cloud computing in genome informatics

Lincoln D Stein\*

DNA  
Sequencing (bp/\$)



# About DNA and computers

- We hit the \$1000 genome\* in ~2016
  - Need to think about the \$100 genome
- The doubling time of sequencing has been ~5-6 months.
- The doubling time of storage and network bandwidth is ~12 months.
- The doubling time of CPU speed is ~18 months.
- The cost of sequencing a base pair will eventually equal the cost of storing a base pair

# What is the general biomedical scientist to do?

- Lots of data
- Poor IT infrastructure in many labs
- Where do they go?
- Write more grants?
- Get bigger hardware?



# Cloud computing providers

- Amazon AWS
  - <https://aws.amazon.com/>
- Google cloud
  - <https://cloud.google.com/>
- Digital ocean
  - <https://www.digitalocean.com/>
- Microsoft Azure
  - <https://azure.microsoft.com/en-us/>
- More...

# Amazon Web Services (AWS)

- Infinite storage (scalable): S3 (simple storage service)
- Compute per hour: EC2 (elastic cloud computing)
- Ready when you are High Performance Computing
- Multiple football fields of HPC throughout the world
- HPC are expanded at one container at a time:



# Some of the challenges of cloud computing:

- Not cheap!
- Getting files to and from there
- Not the best solution for everybody
- Standardization
- PHI: personal health information & security concerns
- In the USA: HIPAA act, PSQIA act, HITECH act, Patriot act, CLIA and CAP programs, etc.
  - <http://www.biostars.org/p/70204/>

# Some of the advantages of cloud computing:

- We received a grant from Amazon, so supported by 'AWS in Education grant award'.
- There are better ways of transferring large files, and now AWS makes it free to upload files.
- A number of datasets exist on AWS (e.g. 1000 genome data).
- Many useful bioinformatics AMI's (Amazon Machine Images) exist on AWS: e.g. cloudbiolinux & CloudMan (Galaxy) – now one for this course!
- Many flavors of cloud available, not just AWS

# Key AWS concepts and terminology

- AWS - Amazon Web Services. A collection of cloud computing services provided by Amazon.
- EC2 - Elastic Compute. An AWS service that allows you to configure and rent computers to meet your compute needs on an as needed basis.
- EBS - Elastic Block Storage. A data storage solution that allows you to rent disk storage and associate that storage with your compute resources. EBS volumes are generally backed by SSD devices.

# Key AWS concepts and terminology

- S3 - Simple storage service. Cheaper than EBS and allows for storage of larger amounts of data with some drawbacks compared to EBS. S3 volumes store data as objects that are accessed by an API or command line interface or other application designed to work with S3. EBS volumes on the other hand can be mounted as if they were a local disk drive associated with the Instance.
- SSD - Solid state drive. A particular type of storage hardware that is generally faster and more expensive than traditional hard drives.

# Key AWS concepts and terminology

- HDD - Hard disk drive. A particular type of storage hardware that is generally cheaper and larger but slower than SSD. HDD drives are traditional hard drives that access data on a spinning magnetic disk.
- Ephemeral storage - Also known as Instance Store storage. Data storage associated with an EC2 instance that is local to the host computer. This storage does not persist when the instance is stopped or terminated. In other words, anything you store in this way will be lost if the system is stopped or terminated. Instance store volumes may be backed by SSD or HDD devices.

# What is a Region?

- An AWS Region is set of compute resources that Amazon maintains (like the Data Center image shown before)
- Each Region corresponds to a physical warehouse of compute hardware (computers, storage, networking, etc.).
- At the time of writing there are 14 regions: (US East (N.Virginia), US East (Ohio), US West (Oregon), US West (N. California), Canada (Central), EU (Ireland), EU (Frankfurt), EU (London), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Seoul), Asia Pacific (Tokyo), Asia Pacific (Mumbai) and South America (Sao Paulo).
- When you are logged into the AWS EC2 console you are always operating in one of these regions.



# What is a Region?

- Current region shown in the upper right corner of console
- It is important to pay attention to what region you are using for several reasons.
  - When you create an EC2 instance (EBS volume, etc) in one region you won't see it in another region.
  - The cost to use many AWS resources varies by region.
  - The region may influence network performance when you are accessing the instance, especially if you need to transfer large amounts of data in or out.
  - Billing is tracked separately for each region
  - Generally you should choose a region that is close to you or your users. But cost is also a consideration.

# What is difference between the 'Start', 'Stop', 'Reboot', and 'Terminate' (Instance States)?

- Start – turn on an EC2 instance that you have previously created
- Stop – turn off an EC2 instance that you have previously created
- Reboot – restart an EC2 instance
- Terminate – permanently stop and destroy an EC2 instance. Any associated EBS volumes may also be destroyed at this time depending on configuration

# What is an AMI/snapshot?

- AMI (Amazon Machine Image) – a template that specifies how to launch EC2 instances
  - Root volume with operating system (OS), pre-installed applications, etc
  - Launch permissions that determine who can use the AMI
  - Specification of (data) volumes to attach when launched
- You can create an AMI for any instance you have created/configured
- AMI can be made public for sharing (region-specific)
- Creating an AMI involves creating a snapshot of the root and any attached volumes. You will be charged to store this snapshot.

# I can not log into my EC2 instance, what might have gone wrong?

- Is your instance running?
- Are you providing the correct path to your key file?
- Is it the correct key file?
- Have you set the permissions for your key file correctly?
- Did you specify a valid user for your AMI (e.g., ubuntu)?
- Did you specify the correct IP address?
- Does the Security Group for the instance allow access for your connection protocol (e.g., SSH) and location?

# How much does it cost to use AWS EC2 resources?

Linux RHEL SLES Windows Windows with SQL Standard Windows with SQL Web

Windows with SQL Enterprise

Region: US West (Oregon)

	vCPU	ECU	Memory (GiB)	Instance Storage (GB)	Linux/UNIX Usage
<b>General Purpose - Current Generation</b>					
t2.nano	1	Variable	0.5	EBS Only	\$0.0058 per Hour
t2.micro	1	Variable	1	EBS Only	\$0.0116 per Hour
t2.small	1	Variable	2	EBS Only	\$0.023 per Hour
t2.medium	2	Variable	4	EBS Only	\$0.0464 per Hour
t2.large	2	Variable	8	EBS Only	\$0.0928 per Hour
t2.xlarge	4	Variable	16	EBS Only	\$0.1856 per Hour
t2.2xlarge	8	Variable	32	EBS Only	\$0.3712 per Hour
m4.large	2	6.5	8	EBS Only	\$0.1 per Hour
m4.xlarge	4	13	16	EBS Only	\$0.2 per Hour
m4.2xlarge	8	26	32	EBS Only	\$0.4 per Hour

Data transfer (GB): In: free or \$0.01; Out: free, \$0.01 or \$0.02

EBS storage (GB/Month): \$0.10

S3 storage (GB/Month): \$0.023 standard, \$0.0125 infrequent access, or \$0.004 glacier

# Why am I still getting a monthly bill?

- Generally you get an accounting of usage and cost on a 30 day cycle
  - Pricing is per instance-hour (now instance-second!) consumed for each instance type.
  - Also charges for storage, transfers, etc
- Be aware of regions!
- Even when an instance is stopped, storage for root or other EBS volumes persist
- Creating AMIs/snapshots requires storage
- Explore the billing and cost management tools of AWS to track your spending, set warnings, etc

# Amazon AWS documentation

[https://github.com/griffithlab/rnaseq\\_tutorial/wiki/Intro-to-AWS-Cloud-Computing](https://github.com/griffithlab/rnaseq_tutorial/wiki/Intro-to-AWS-Cloud-Computing)

<http://aws.amazon.com/console/>

# In this workshop:

- Some tools (data) are
  - on your computer
  - on the web
  - on the cloud.
- You will become efficient at traversing these various spaces, and finding resources you need, and using what is best for you.
- There are different ways of using the cloud:
  1. Command line (like your own very powerful Unix box)
  2. With a web-browser (e.g. Galaxy): not in this workshop



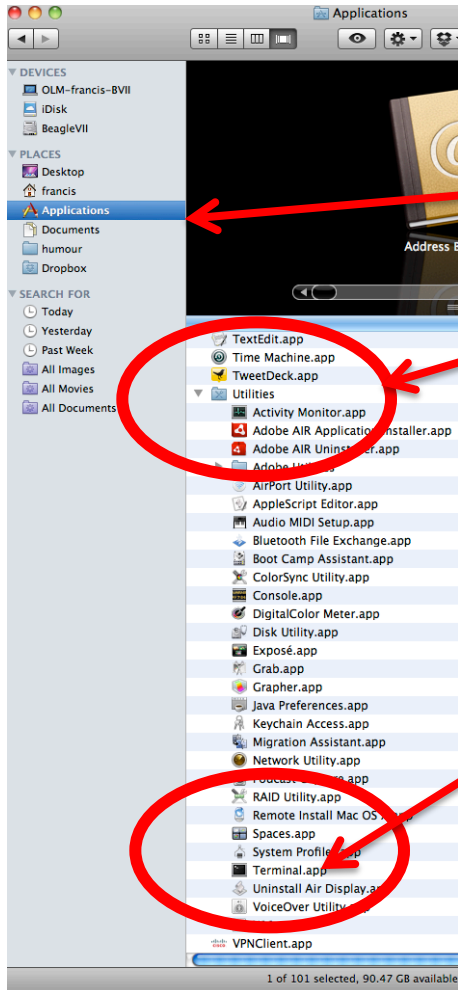
# Things we have set up:

- Loaded data files to a web server
- We brought up an Ubuntu (Linux) instance, and loaded a whole bunch of software for NGS analysis.
- We will clone this and create separate instances for everybody in the class.
- We've simplified the security: you basically all have the same login and file access, and opened ports. In your own world you would be more secure.

# Logging into Amazon AWS

# Opening a 'terminal session' on a Mac

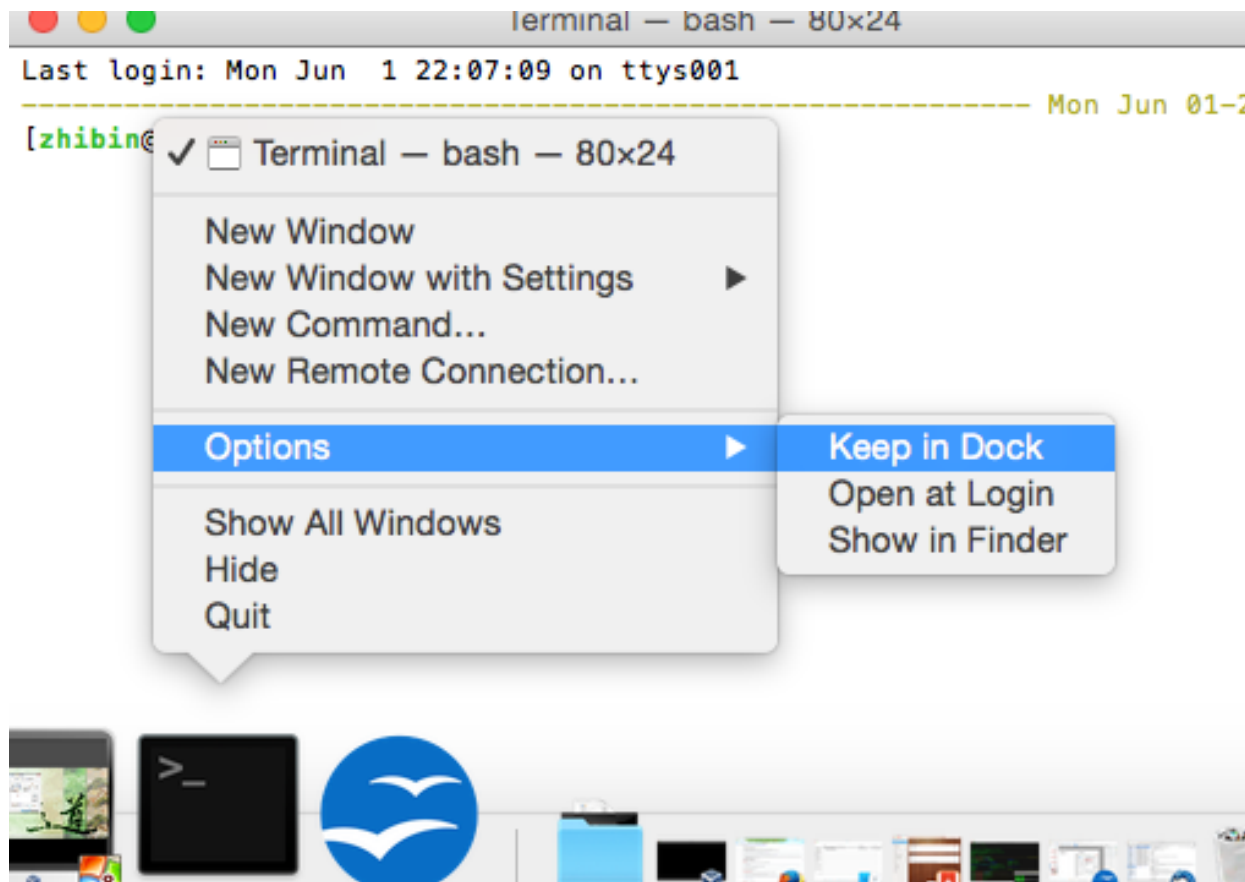
In a Finder window  
'Applications' -> 'Utilities' -> 'Terminal'



Or on your dock



# Add the terminal App to your dock



# Creating a working directory on your Mac called 'cbw'

```
obis-air:~ ogriffit$ pwd
/Users/ogriffit
obis-air:~ ogriffit$ ls
Applications      Desktop           Dropbox           Movies            Public            gittemp          temp
Attachments       Documents        Google Drive     Music             bin               igv              ncbi
Box Sync          Downloads        Library           Pictures          git
obis-air:~ ogriffit$ mkdir cshl
obis-air:~ ogriffit$ cd cshl
obis-air:cshl ogriffit$ ls -la
total 0
drwxr-xr-x  2 ogriffit  staff   68 Nov 13 22:18 .
drwxr-xr-x+ 58 ogriffit  staff  1972 Nov 13 22:18 ..
obis-air:cshl ogriffit$
```

mkdir cbw  
cd cbw

# Go to course wiki, “Accessing the cloud” page

## Download the certificate

Day 1 and Day 2 refer Home in the Welcome seq wiki Pre-workshop Materials Day 1 Day 2 Day 3

### Welcome

*Ann Meyer*

### Introduction to Cloud Computing

*Obi Griffith*

#### Lecture

- We have set up 30 instances on the Amazon cloud - one for each student. In order to log in to your instance, you will need a security certificate. If you plan on using Linux or Mac OS X, please download this certificate. Otherwise if you plan on using Windows (with Putty and Winscp), please download this certificate.
- Detail instructions can be found here.

Module 1: Introduction to RNA Sequencing Analysis

[https://bioinformaticsdotca.github.io/RNAseq\\_2019](https://bioinformaticsdotca.github.io/RNAseq_2019)

# Viewing the 'key' file once downloaded

```
cat CBWNY.pem
```

```
obis-air:cschl ogriffit$ cd ~/cschl/
obis-air:cschl ogriffit$ ls -la
total 8
drwxr-xr-x  3 ogriffit  staff   102 Nov 13 22:21 .
drwxr-xr-x+ 58 ogriffit  staff  1972 Nov 13 22:18 ..
-rw-r-----@ 1 ogriffit  staff  1696 Nov 13 22:21 CSHL.pem
obis-air:cschl ogriffit$ cat CSHL.pem
-----BEGIN RSA PRIVATE KEY-----
MIIEPgIBAAKCAQEAJ5gwmTby9QZ2Idz+ugiEQQHW6Ps0ZAZFvr+mWdN4pKpccaVmDh7XjceOLF
OkJzaP9+jj0kSF0yNinitoB32DgrmVhgNhyheEqH5XMn28szxUj1EuoNXAogNuY7mWmo6MoWssSW
Rqy+rj19vMGQn5rsnMLjCM1smebPoqYOL8EPa1ccRbdGXG1dMTLCC1ho/Hk9bZweamGiZLaAWVmF
zOK/L0zXgY3K4cwaL48HV6oGuMh5lTDpnobxXghQ4oC5Mej+DpCRF8C+EG2uNDuyulzRjFQmFBV2
GKDWDwhdgGmKmX9IpMT9ubvNoQPy0vYlvM80eG3cMbz2IZpaNryihwIDAQABAoIBAQCZYT0TvF04
a3DdCEEC/rN9HMaS+bjFkm0kp9RTi15XJhTPvBmptjzibA6gWJfDaXgKIQGbzXJrEkxwCR2IB03v
0LV7jEcomZ2ggRMDPeJitFoUCuDnkZZtivppSk2az0zeaD+0/ZeqPx0L+Yr+7HSbpVLVoxEV/l5a
xDuCawBMSY2cnGWKfEBLSPnB6fGZj8luGzv0aP/CETx/K78TIS56m4yrTIQIeEPfFt/PQr/EUqoL
7co5oy9K3sD1noPLDhk3vJa1VNrMjHkMZLkbZuaoHPzgsQHninm80Ca25WWTGsSZ8vQsBIUTLGI1
W7lzXH3wD1jJNd+03QK4bnKaZ+DZAoGBAPVpisa49JY/6K2f9B8naqtX/ljzVWTl3Q7r6t6uh21Y
oexmC8eJ2wQwd0qNjZWVxSMVksIwdM6xcsBIJRMmltWTVdmD0fkDv0fjd8CM4nctH76tvSvZz02e
qI9wSshHY1fh+09CoLZeefFSURxqWbkJfREjoZ4UGUWmi3k1rxC9AoGBAMTB1BB0WQ+5ojzQYu0L
Q4YrsIPg1/ni0WmJ+05vcTCJ2aeI88VhK5c2PoXPWwiJ9CdD2VFZDiCm2XuJA5iwJmnhuwGGHHEn
BFBqEF/ueJrW+r43pRcYRuRiXjiH4mQQlK4Zemecym5fAHvxZxq4fs2kWfMPySFaVufcP0VC7X6T
AoGBAMhro0xbrFQwaU0yh9oRhMneGPhn8WtvVjNjc/LcMfmZEtRPGnuhF965/hJCvEhXgiH+8LXo
4NwUixSBvtXnA/P0WX5Ea2ykIth2Kkx0Qlb14SEGHqH7RZ0saRiLqmcZ9gXFpkm6rimByrDMezVr
nU7CcwNWSB0ja0gluZoJv6k5AoGBAJJufsmD5ZhkaS+lTtpnlZtXDIk5XsMkYQGQpS0clzqufQPI
UtPEm3Jv9lwTktDQSpqmTifShUcbpaPgtoJ+JjiKvGhH7QbxKK7II00kULG760SD+S0U972Rdj3Q
M1aRWHWxlH1kH0vDXFLhuAAU6poVBLR2PRPLbf4k1hmv05xtAoGBAJVQy1GF8uVNwk0CNzLIqmkY
uk9M24hfqn3N2GY3Zgqf43bD4kdYgL4rvsgp08QzotPf+19kvlCv0ciolsjEHLyUdlyPGzj4CTTH
1f1RoGHmYzVn9VuFTu4hJ17J+uwgXgIr9Sx/UTjwkmCjPf7CEyIuGxaThG/ZoR9stufZB5db
-----END RSA PRIVATE KEY-----obis-air:cschl ogriffit$
```



# Changing file permissions of your 'key' file (Mac/Linux)

## ls -l (long listing)

```
drwx-----+ 67 ogriffit staff 2278 22 May 21:25 ../  
-rw-r--r--@ 1 ogriffit staff 1696 22 May 21:31 CBWNY.pem
```

rwX : owner

rwX : group

rwX: world

r read (4)

w write (2)

x execute (1)

Which ever way you add these 3 numbers, you know which integers were used (6 is always 4+2, 5 is 4+1, 4 is by itself, 0 is none of them etc ...)

So, when you have:

**chmod 600 <file name>**

It is "r" for the the file owner **only**

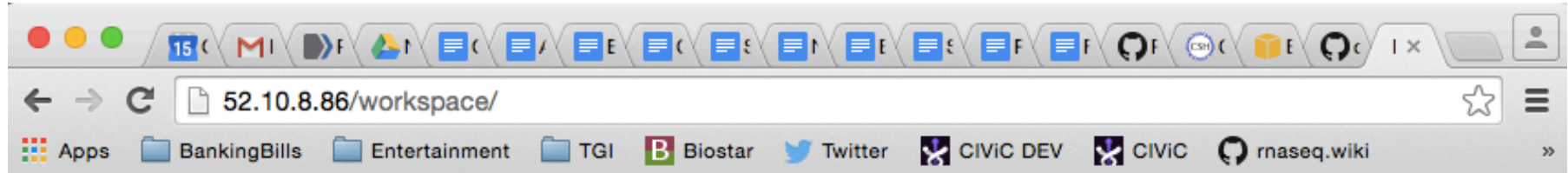


# Logging into your instance



## Mac/Linux

```
cd ~/cbw  
chmod 600 CBWNY.pem  
ssh -i CBWNY.pem ubuntu@[YOUR PUBLIC IP]
```

# Copying files from AWS to your computer (using a web browser)



## Index of /workspace

<a href="#">Name</a>	<a href="#">Last modified</a>	<a href="#">Size</a>	<a href="#">Description</a>
 <a href="#">Parent Directory</a>		-	
 <a href="#">Homo sapiens/</a>	2015-11-13 06:45	-	
 <a href="#">README.txt</a>	2014-06-17 23:53	5.3K	
 <a href="#">bam-demo/</a>	2015-11-14 21:03	-	
 <a href="#">data/</a>	2015-11-13 01:39	-	
 <a href="#">scratch/</a>	2015-11-13 19:43	-	
 <a href="#">tools/</a>	2015-11-13 01:54	-	

*Apache/2.4.7 (Ubuntu) Server at 52.10.8.86 Port 80*

`http://[YOUR PUBLIC DNS OR IP]/`

# Logging out of your instance

**Mac/Linux – simply type exit**

exit

Note, this disconnects the terminal session (ssh connection) to your cloud instance. But, your cloud instance is still running! See next slide for how to stop your instance.

# So, at this point:

- Your Mac desktop is ready for the workshop
- If it is not, you know where to get the information you need
- You know how to login to AWS
- The next step is to login to your linux machine on AWS and learn the basics of a linux command line

**We are on a Coffee Break &  
Networking Session**